加微信：642945106 发送“ 赠送 ”领取赠送精品课程
发数字“ 2 ”获取众筹列表

34 | 服务需求控制管理：每种需求都是必需的吗？

2020-02-12 庄振运

性能工程高手课 进入课程 >

讲述：庄振运
时长 16:13 大小 13.01M

你好，我是庄振运。

上一讲，我们探讨了如何通过提高互联网服务的效率，降低对公司服务容量的要求。今天我们讨论另一个有效手段——互联网服务的内部需求控制管理。

互联网公司内部，往往有很多后端服务，比如 Key-Value，也就是键值数据库服务。公司内部对这些后端服务，会有很多使用的需求。需求自然有合理的，也有不是很合理的。我们要做的，就是确保那些合理的需求，能够使用这些后端服务，并且将那些不合理的需求量挡回去，或者，让提出需求的内部客户提高容量使用效率。

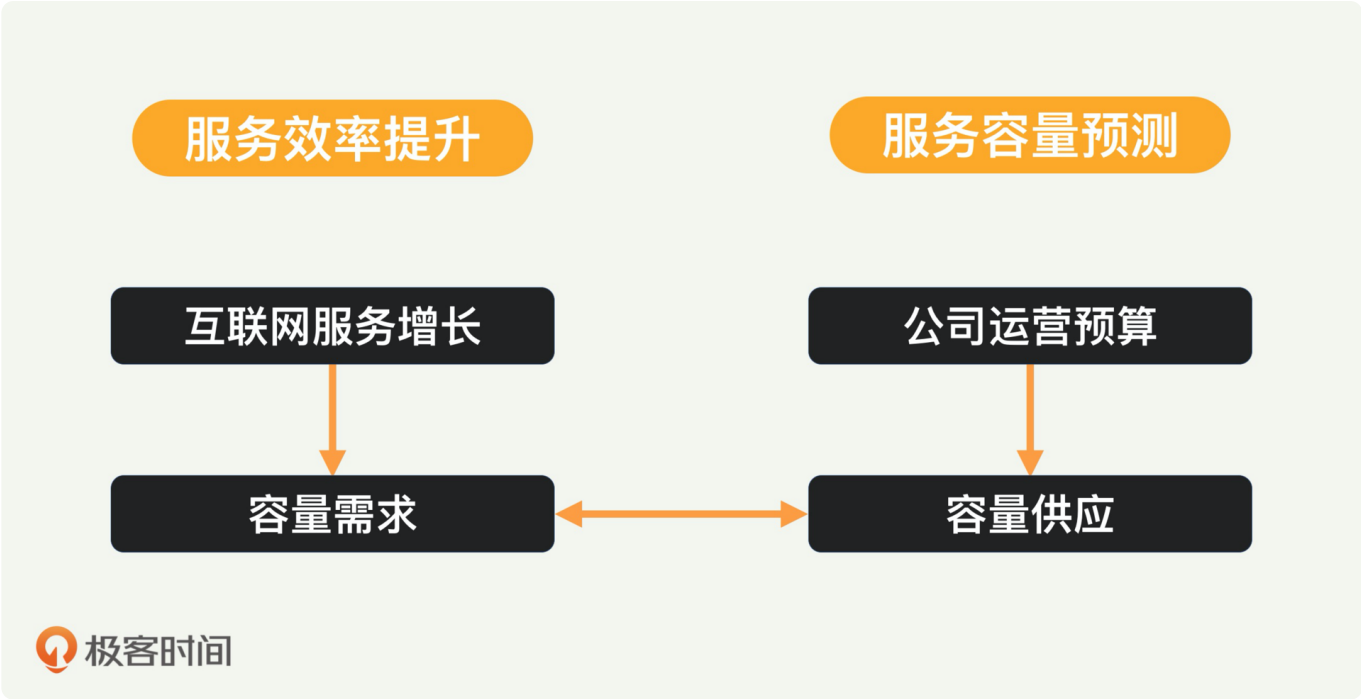
你不要小看了互联网服务的需求管理，它对保证容量供给和服务质量很重要。因为公司给一个服务的容量总是有限的，如果不能控制需求，那么有限的容量很快就会被用光，从而影响公司的正常运行。

今天我先讲讲服务需求控制管理的总体目标，然后分享一整套大规模需求控制管理的方法给你。这套方法是我在多年实践中总结出来的，实践证明它是行之有效的。

为什么要做服务需求控制？

想做好大规模需求控制的管理，你就需要先宏观地弄清楚服务需求控制的目标。

一个公司要想成功、有效地运行互联网业务，它就应该尽量让各个内部服务的容量需求和供应尽可能接近。



如上图所示，左边是公司业务增长，导致容量需求越来越大；右边是公司预算成本有限，容量供应有限。

我们总是希望这两者能够匹配，越接近越好。如果容量供应不能满足服务需求，那么会导致部分业务的发展受到阻碍。如果容量供应过量，就会有闲置容量，造成浪费，从而降低公司基础架构的效率。

要降低容量需求有两种方法，一种就是第 33 讲我们谈过的**提高服务效率**，另外一种就是这一讲要说的**控制需求**。这二者最好互相配合，共同努力。

需求控制管理难在哪里？

现在我们就来看看需求控制管理要怎么做。不过，要理解这套方法为什么是这样的，又为什么可以说它是“行之有效”的，你需要先了解做需求控制有哪些挑战。我总结一下，主要有下面几个方面的挑战：

一是很多公司，尤其是大公司，**服务的规模会越来越大**。这个规模表现在很多方面，包括服务数量很多，使用服务的客户也很多。注意这里的客户不仅仅是外部客户，更是内部客户。这就要求我们的工作不能仅靠一个人或者一个团队，必须是分布式的。

二是服务的内部**客户的容量使用情况很不一样**。有的客户对容量使用量很小，增长也不快；有的使用量就很大，或者飞速增长。这就要求我们，在宏观上需要对客户的使用设置一定的配额，否则一个服务的容量，一旦被某些客户用光，其他客户的性能就会非常差。

三是内部**客户的资源使用可能不稳定**。客户使用的不稳定，会导致特别大的峰值出现，对服务的稳定性造成影响。所以，我们必须在微观上，实时监测客户的资源使用情况，一旦发现超出期望的效率衰退，就需要及时通知客户，让他们尽快解决。

四是**停止服务的代价高**。一个内部客户一旦开始使用一个服务，就很难让人家离开。所以必须严把入口关，不要随意地让服务上马。

五是资源使用效率的提升，必须**要经常与用户沟通**，和用户一起不断地提高使用效率，同时也需要开发一些工具来帮助用户。

需求控制框架如何构成？

我把整个需求控制的框架按照逻辑，分成了五大模块。如下图所示，我们一个一个地来展开阐述。

服务入口控制：需求审核

第一个模块是服务入口控制。一个服务的需求在开始使用服务之前，必须进行上马前的审核。这个审核，需要统筹考虑需求优先级、和什么产品相关、对公司业务的影响、需求量的大小和轻重缓急、是不是已经计划好的等很多因素。

每个服务需求的请求，要求以任务的形式进行，审核人员要定期的处理这些任务。为了帮助审核人员做出适当的决定，请求者最好提供足够的信息。一般来说，按照刚刚提到的几个方面来准备就可以，越详细越好。

出于平衡利益的考量，审核人员的组成十分多样，至少要包括服务开发团队、运维团队，以及提供容量的团队。为什么这样设置呢？给你举个例子，服务的开发团队，可能希望服务的内部客户越多越好，因为客户多的话，这个服务就变得越来越重要。服务的运维团队，可能更关心服务和系统的稳定性，不要出现各种性能和可靠性问题。而提供容量的团队为了保证容量供应，可能不希望用户有太多的需求。

经过审核后，每个请求会产生几种不同的结果：或批准，或拒绝，或需要客户提供更多信息，或者要求客户降低请求的大小，又或者推迟这个请求。

之所以把这个审核放在第一位，是因为我在实践中发现，很多的客户请求都是“拍脑袋”想出来的，其实根本不合理，而且请求的大小，往往是狮子大开口。当你审核一个“需要几千

台服务器容量”的请求后，你必然要与提出者沟通。而在很多情况下，他们的请求在沟通后降到了只需要几台服务器。

这个“沟通”也不难，你只要对每个请求多问几句，比如这三个“劝退”问题：

> 你为什么有这么大的需求？
>
> 你能讲讲你的理由吗？
>
> 能不能推迟一下这个请求？

差不多有一半客户在答完这三个问题后会降低请求的大小，甚至很多会直接取消请求，说：“仔细考虑后，我们其实不需要这个请求。”

分布式的需求控制管理和效率提升

第二个模块是分布式的需求控制管理和效率提升。

为了适应大规模服务的场景，需求控制的方法必须是分布式的。我们一般是把整个公司的各种使用，分成差不多 10 到 15 个产品组，比如按照产品种类来分。这样我们就可以针对这些产品组，分别合作，让每个产品组设置特定人员和团队来和我们合作。

拼课微信：1716143665

为什么这么做呢？因为只有这样，我们的工作才能合理扩展。

我曾经为一个互联网服务做过三年的需求控制，学到了很多经验和教训。这个服务有几千个内部客户，如果只有我一个人和所有的客户打交道，根本行不通，累死也做不过来。

更重要的是，每个产品组的内部人员，其实才是最适合审核本组需求的人，因为他们对这些请求最清楚。这个请求到底需不需要？请求的大小到底合不合适？由他们来做审核，远远比我做得好。

产品组级别的审核，本质上是一种有效的，可扩展的分布式模型。而且，除了需求控制和审核，还有其他一些类型的工作，也更适合在产品组级别执行。比如下面我们会提到的配额限制等。

采用这个模式，所有的服务需求，都会先确定是哪个产品组，然后将需求任务分配给产品组，让他们进行产品组级别的审核。产品组的审核决定也不是随意决定的，同样要基于几个因素，包括请求的优先级（例如功能的重要性）、增长配额以及产品组内下层团队之间的协作等。

宏观需求增长控制：设置增长配额

第三个模块是在宏观上控制增长，也就是设置增长限额和配额。我通过四个问题来阐述。

1. 在哪个级别设置增长配额?

我刚刚讲了，每个产品组会对自己内部的服务需求进行审核。如果没有产品组级别的配额限制，产品组有可能胡乱批准他们自己的请求，从而滥用这个服务。如果对他们的服务需求，加上配额限制，事情就好办了。所以，**增长配额必须设置在产品组的级别**。

另外，在产品组下面的级别，也可以设置配额。因为有的产品组很大，内部也有几十甚至几百个团队，所以每个团队也需要设置相关的增长配额。这也有利于产品组内部的协调。这里的要求就是，产品组内部的配额总和，必须不能超过产品组级别的总配额。

2. 怎样设置增长配额的频率?

这个设置可以是一年、半年，或者是一个季度。太频繁会增加工作负担，但是时间太长也有弊端。所以，我们一般都是半年或者一年设置一次。如果是一年，一般是年初设定一年的增长配额。

3. 如何设置增长配额的大小?

每个产品组的实际配额大小也需要考虑几个因素，比如产品的特征、重要性、预期增长、产品发布计划、内部客户端效率等等。对每个产品组的情况都了解后，就可以根据实际情况和业务影响，来平衡各个产品组的配额。当然前提是，总体配额不超过总容量。

4. 如何及时地控制需求的增长，保证不超过配额?

我们需要**搭建一些工具和配额使用情况面板**。这样就可以每天监控每个产品组的资源使用情况，并将使用情况与增长配额进行比较。如果产品组的资源使用量，连续 7 天比每日的配

额快，我们就通知产品组，并创建对应的任务。

任务的优先级别在一开始要设置成最低的。但是如果产品组的使用量，已经超过了年度总增长配额，那么你可以将任务更改为高优先级来解决。如果有连续 3 周以上的实际使用量，都超过了年度总增长配额，并且产品组没有采取任何措施，那么对应的任务就会提升优先级到最高级。

微观增长控制：资源使用的实时监测

第四个模块是在微观上进行增长控制，就是实时地进行资源使用检测。

产品组对一个互联网服务的使用有很多实例。虽然每个实例的情况都不同，需要区别对待，但不会变的核心点是：每个使用服务的实例，都需要**进行实时的、微观层面的监测**，以确保它的效率不衰退。

一旦你发现实例的效率衰退，就需要及时地通知这个实例的所有者。一般是以任务的形式发给客户，让客户诊断和根因，然后解决。

这个部分，和刚刚讲过的宏观层面的配额限制是互相呼应的。只有在微观增长方面检测和控制得好，才能更容易保证配额不超标。

资源使用效率：不断提升

最后一个模块，就是不断提升资源使用效率。最好的措施，就是我们上一讲说过的提高服务效率。这里的服务效率，主要是客户端的效率，目的是满足内部客户基本要求的情况下，尽可能地降低资源量。

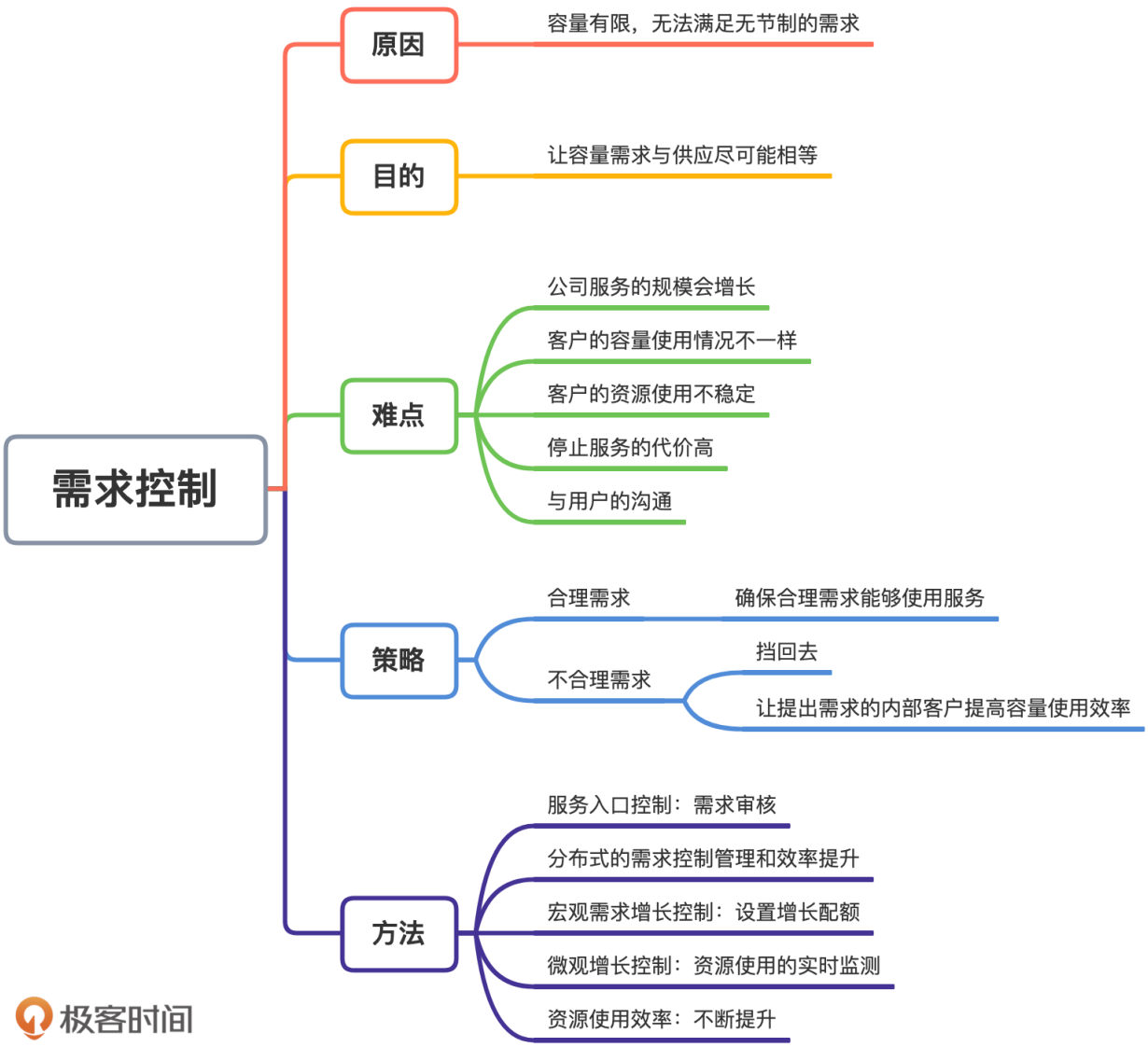
但我们实际中碰到的一个难点是，内部客户通常只对自己的指标比较了解，比如 QPS，或者数据的大小。既然我们的最终目的是节省成本，那么，就需要让客户直观地了解指标和成本的对应关系，比如 1TB 的数据相当于多少钱。

有了指标和成本的对应关系，产品组就能方便地按照任务的重要性来安排工作、调配人员。一个能节省 1 千万元的任务，当然比只能节省 1 百万元的任务更重要。

为此，我们针对每个互联网服务，都提出了一种易于计算和理解的对应指标。比如，对某个存储服务，我们就考虑了数据大小和 QPS 这两个输入指标，然后根据整个服务的成本来做映射。一般来讲，为了反映最新的服务状态，这个指标是需要定期更新的，例如，服务增长、硬件效率、服务效率等。

总结

这一讲我们讨论了容量控制的管理方法，这套方法可以有效地降低服务的容量需求，也就节省了公司的运营成本。实际执行这条管理方法的核心是：在不影响公司正常发展和内部客户合理需求的条件下，尽量降低内部客户对每个互联网服务的需求。



总体来讲，这套方法的核心是分布式和系统性。对于新的内部客户需求，要进行审核以确保需求请求是合理的。而对于现有客户的需求，要在宏观级别，确保需求配额不要超过；在微

观层面上，确保资源使用效率不退化，还要不断地寻找效率提升优化的机会。

唐代诗人岑参有首很有名的边塞诗，有两句是描写塞外的寒冷：“将军角弓不得控，都护铁衣冷难着。”说的是边塞都护府的将军，手冻得拉不开弓，几乎不能控制弓箭了，铁甲也冰冷得让人难以穿着，非常辛苦和困难。我们做容量控制管理的工作也着实不容易，需要和各个内部客户产品组打交道，互相理解，为了共同的目标和公司的利益一起工作。

思考题

你的公司里最大的互联网服务是什么？这个服务的容量增长是怎么控制的？

如果这个服务的客户要求实现一个新的功能，所以需要很多服务器，而你负责这个服务的容量管理，你会如何回答客户呢？

欢迎你在留言区分享自己的思考，与我和其他同学一起讨论，也欢迎你把文章分享给自己的朋友。

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　33 | 服务效率提升：如何降低公司运营成本？

下一篇

精选留言 (1)

写留言

夜空中最亮的星（华仔...

2020-02-13

老师真是 文武双全 佩服

展开

作者回复: 谢谢!